Weekly Report
10/20/2014 - 10/26/2014

Jing Xia

October 26, 2014

1 Summary

This week I mainly focused on the rank visualization project. A new version of top-10000 data is collected from the Wikipedia data dumps.

2 Projects

2.1 Project 1 - Rank Visualization

2.1.1 Data Preparation

In the past week, we've collected almost 12 months of the raw wikipedia dumps data (6 months of 2011 and 6 months of 2014). The 2011 dataset is used for comparisons of the current dataset, and the 2014 dataset is the latest update.
After collecting the raw datasets, we also need to make a statistical count of everyday page view, and after that taking the top-10000 of the everyday page view. Wikipedia Special pages and Index page are eliminated from the dataset.
Based on the Top-10000 dataset, we modify the codes to satisfy a discrimination of total items (K) and total items in display (TOTAL).
Problems occurs with loading large times series dataset. The processing time is too slow. But the visualization part actually require only K items per day instead of TOTAL items. TOTAL items are required only after user selection and only a very limited amount of them. A time series database is in consideration of organizing such dataset.

2.2 Project 2 - Data Inspection

I had a pre-leaving meeting with Dr. Ebert about the two projects. He mentioned two concepts that might be related to this project: *visual saliency* and *data provenance*. He also mentioned some related projects that I can refer to.

3 Paper Reading

-

4 Miscellaneous

-

5 To Do List

1. Download and preprocess Wikipedia dataset.
2. Finalize the system design and prepare for paper writing.
3. Coming home, finally.

References